基于支持向量机的中文极短文本分类模型 *

王　杨†，许闪闪，李　昌，艾世成，张卫东，甄　磊，孟 丹
(安徽师范大学 计算机与信息学院, 安徽 芜湖 241000)

摘　要：随着智能终端设备的不断普及，微信、网络即时新闻、电商客户产品评论等富含极短文本数据的信息呈爆发式增长。为了有效提取极短文本中的关键特征信息，提出了一种基于支持向量机的极短文本分类模型。首先对原数据进行数据清洗并利用 Jieba 分词将清洗过的数据进行处理；再将处理后的数据存入数据库，通过 TF-IDF 进行文本特征的提取；同时，利用支持向量机对极短文本进行分类。经过（1-0）检验，验证了模型的有效性。实验以芜湖市社管平台中的 9906 条极短文本数据作为样本进行算法检验与分析。结果表明在分类准确率方面，该方法相比于朴素贝叶斯、逻辑回归、决策树等传统方法得到有效提高；在误分度与精确度指标上匹配结果更加均衡。
关键词：支持向量机；Jieba 分词；极短文本分类；TF-IDF
中图分类号：TP391.1　　doi: 10.19734/j.issn.1001-3695.2018.06.0514

Classification model based on support vector machine for Chinese extremely short text

Wang Yang†, Xu Shanshan, Li Chang, Ai Shicheng, Zhang Weidong, Zhen Lei, Meng Dan
(*School of Information & Computer Science, Anhui Normal University, Wuhu* Anhui 241000, *China*)

Abstract: With the increasing popularization of intelligent terminal devices, information containing abundant extremely short text data, such as WeChat messages, online instant news and customers' comments on e-commerce websites have been experiencing explosive growth. In order to effectively extract the key features from the extremely short texts, this paper proposes an extremely short text classification model based on SVM. Firstly, by the data cleansing on the original data, the cleaned data is processed by the Jieba segmentation and TF-IDF. Then the (1-0) test verified the validity of the model. Finally, 9906 pieces of extremely short texts in Wuhu City Community management platform are used as the sample in this experiment. The results show that the proposed method can effectively improve classification accuracy compared to other traditional methods, such as Naive Bayes, Logistic regression and Decision tree. At the same time, the matching results in terms of misclassification and accuracy are more balanced.
Key words: support vector machine; Jieba segmentation; extremely short text; TF-ID

0　引言

随着各种智能终端和社交软件的广泛应用，用户针对社会热点、政府行为评判的表达方式更加广泛、多元、便捷。其中各种各样的评判较多采用不完整形式的极短文本加以表达。如何从不完整的极短文本中快速提取出有价值的信息，对决策者显得极为重要。着眼于当前自媒体蓬勃发展的大数据时代，人们更加习惯于通过 Twitter、Facebook、微博、微信等在线社交平台，以简短精炼的朋友圈动态，寥寥数字的问题反馈等形式传递情感、表达诉求。这种文本形式具有碎片化、即时性的特征，因此传统的文本分类方法就难以快速提取此类文本中的信息。本文提出了一种基于支持向量机的极短文本分类模型。

现有的文本分类方法主要有以下两种：a)聚类词嵌入法，该方法将一个 k-均值算法应用到文档的单词向量上，以获得一个固定大小的集群集合。每个文本被表示为一个超级单词嵌入包，计算每个超级单词嵌入在各自文本中的频率，即可得出文本分类[1]；b)频率加权法,将所缺少的条款计算在内，计算出现有条款的权重，结合 SVM 分类器，得出最优分类性能[2]。

在对文本分类的研究中，支持向量机得到了广泛应用。目前基于 SVM 的文本分类技术主要有以下几种：a）改进混合核函数分类方法[3],通过将学习能力较强的核函数与泛化能力较好的核函数重组为混合核函数，达到提高分类效果的目的；b）基于增量学习的 SVM 分类方法[4],充分考虑新增样本对初始样本的影响，引入边界支持向量，提出基于边界支持向量的增量学习算法，在训练速度和训练精度方面有一定提高;c）特征选择分类模型[5],针对传统的卡方特征选择方法的局限性，提出新的类内信息优化卡方统计特征选择方法。有效提高了模型的特征选取能力。

1　相关概念及方法

1.1　极短文本

狭义文本是指书面语言的表现形式，从文学角度说，通常是具有完整、系统含义的一个句子或多个句子的组合。广义文本是指任何由书写所固定下来的任何话语。在狭义文本的基础上，文本长度不超过 160 个字符的文本称作短文本[6]，比如通过微博、网易云评论，中文垃圾短信，垃圾邮件等形

收稿日期：2018-06-29；**修回日期：**2018-08-28　　**基金项目：**国家自然科学基金资助项目(61871412)；安徽省自然科学基金资助项目(1808085MF178)；安徽省人文社科基金资助项目（SK2014ZD033，AHSKY2017D42）
作者简介：王杨（1971-），男，教授，博士，主要研究方向为社会网络、机器学习、大数据(wycap@126.com)；许闪闪（1995-）女，硕士研究生，主要研究方向为机器学习；李昌（1998-）男，硕士研究生，主要研究方向为机器学习;艾世成（1994-）男，硕士研究生，主要研究方向为机器学习；甄磊（1995-）男，硕士研究生，主要研究方向为机器学习.

成的文本。他们是当下研究文本分类的主要对象。随着信息技术的发展与生活节奏的加快，出现了一类用更加简洁的文字来描述事物的文本，这就是极短文本（extremely short text, EST）。下面给出极短文本的定义。

定义 1　极短文本是指书面语言的表现形式，可能包含某种客观陈述或者评价建议，不一定具有完整、系统含义，由几个词语或者短语组成的文本，句子长度一般不超过 15 个字。

极短文本主要来源于互联网，具有数量大、噪声强、内容特征极稀疏等特点[7]。生活中诸如共享单车故障的报错描述，淘宝商品的简短评价，全民社管上报案卷等信息都属于极短文本。有效的识别并分类极短文本，达到快速处理极短文本的内容的目的，在数据应用、公司管理、政府决策等方面有着重要的意义。

1.2　增益特征向量

在分析极短文本时，分词并选取特征词对后续研究尤为重要。由于文本过短，从已知内容中，一般仅可以提取出 3~4 个关键词。显然，如果仅基于这些特征词建立下文所述模型，则信息量不足，无法保障结果的精确度。因此，本文提出了一种特征词增量模型。

下面以社会垃圾的信息管理为例说明该模型中特征词的扩容。首先，分析社会管理上报案卷的极短文本并提取特征词，记 $B(B_1,B_2,\cdots,B_u)$ 为特征词组成的特征向量，此处 u 的取值较小，一般不超过 4；其次，进一步分析文本可知，“水上漂浮”“绿化带”“路面”等词语描述了垃圾的位置信息，可以将它们概括为一个新的特征词，记作 B_{u+1}；依此类推，当 $u=5,6,\cdots,m$，便得到增益特征向量 $B(B_1,B_2,\cdots,B_m)$。在 u 取值大于等于 5 时，特征向量就具备了较强的代表性。

1.3　文本预处理

如图 1 所示，文本预处理分三个步骤进行:a）加载原始数据，将其中混杂的文本与其对应的类别号分离开来;b）过滤停用词，考虑到原始数据中口语话较严重，存在大量无意义的停用词，因此需进行去停用词处理,本文所使用的停用词表为《哈工大停用词表》[8];c）利用 Jieba 分词工具对纯文本进行分词。

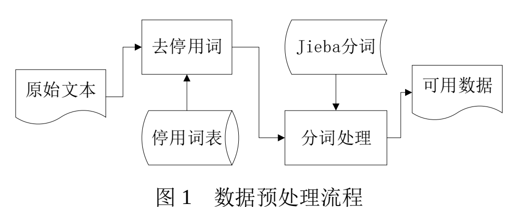


图 1　数据预处理流程

Fig. 1　Data preprocessing process

1.3.1 分词处理

Jieba 分词是一种使用 Python 语言开发的中文分词工具。它有三个主要特点：a）支持三种分词模式：精确模式、全模式、搜索引擎模式；b）支持繁体分词；c）支持自定义词典。Jieba 分词的实现基于以下三个原理：a）基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（Directed Acyclic Graph DAG)；b）采用动态规划查找最大概率路径，找出基于词频的最大切分组合；c）对于未登录词，采用了 Viterbi 算法和基于汉字成词能力的 HMM 模型。

本文将采用 Jieba 分词中的精确模式。该模式是 Jieba 分词中最基础和自然的模式，它试图尽可能精确地划分语句，因此适合极短文本分析。

1.4　TF-IDF 特征提取

预处理文本之后，需要采用 TF-IDF 特征提取法在所得文本中提取关键词并据此进行建模。TF-IDF 用以评估一个字或词对于一个文件集或语料库中某份文件的重要程度。对于特征词 w 而言，它的特征提取函数为：

$$f(w)=TF(w)\times IDF(w)\times \log[N/n(w)+1] \tag{1}$$

其中，TF（特征项频率）是指在一个文本中，某个特征项（可以是字或词）的出现次数与文本中所有特征项的出现总次数的商。如果某个特征项在一篇文本中出现的次数较多，则表明该特征项可能较好地描述了该文本的主要信息，适合用于分类。

而选择 IDF（反文档频率）作为另一个因数的主要思想是：只有少量文本才包含的特征词比大量文本中都包含的特征词要更重要，更加有利于区分文本的类别。若文本集中包含特征词 w 的文本数量越少，则表示 w 的类别区分度越好。IDF 可以减弱在大量文本中都含有的特征词的重要程度，也可以加强只有少量文本包含的特征词的重要程度。

因此，特征项频率 TF 与反文档频率 IDF 经常结合起来使用。常用式（2）计算 IDF：

$$IDF(W)=\log[N/n(w)+1] \tag{2}$$

其中，N 为文本总数，n(w)为包含 w 的文本数。

TF-IDF 特征提取法利用式（1）计算出文本中每个特征词的 TF-IDF 权重值，并对其进行降序排序，然后根据预先设定的筛选条件筛选出满足要求的前 n 个特征词，从而实现了对原特征空间的降维。

1.5　支持向量机

通过上述文本分类方法确定了极短文本的若干个特征词。在支持向量机中，对于待分类的样本，将寻找一个所谓最优超平面，即令样本之间的间隔达到最大，这对于提高 SVM 分类器的泛化能力，增强分类器对于未知样本的预测准确率具有很大帮助[10]。

现以二维线性可分数据为例讨论支持向量机分类器的构建,如图 2 所示。

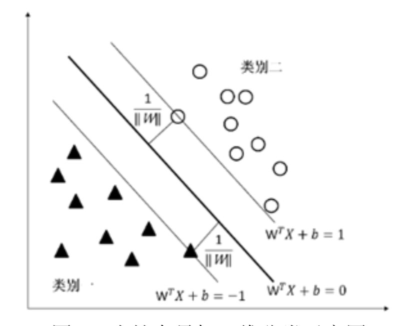


图 2　支持向量机二维分类示意图

Fig. 2　Schematic diagram of two-dimensional classification of SVM

假设现有 P 个线性可分样本 $\{(X^1,d^1),(X^2,d^2),\cdots,(X^P,d^P)\}$，$d^p\in\{-1,1\}$，对于某一输入样本 X^p，期望输出其分类结果 d^p。

定义超平面方程为

$$W^TX+b=0 \tag{3}$$

其中：X 为输入，W 为权值向量，b 为偏置。则任一训练样本都满足：

$$d^p(W^TX^p+b)\ge 1 \tag{4}$$

当等号成立，则样本点分布在超平面附近，称为支持向量。

为寻找最大间隔平面（最优超平面），由解析几何知识，定义样本空间内任一点 X 到最优超平面的距离为

$$r=\frac{W_0^T X+b_0}{\|W_0\|} \tag{5}$$

由式（5），支持向量到超平面的代数距离为

$$r=\pm\frac{1}{\|W_0\|} \tag{6}$$

由上式可知，要找到最优超平面，则只需 $\|W\|$ 最小。此时优化问题可以转换为在式（4）约束下，求：

$$\min\frac{\|W\|^2}{2} \tag{7}$$

引入 Lagrange 函数：

$$L(W,b,\alpha)=\frac{1}{2}W^TW-\sum_{p=1}^{P}\alpha_p[d^p(W^TX^p+b)-1] \tag{8}$$

则此时问题转换为求 Lagrange 函数的最小值。对 W 和 b 分别求偏导，并使结果为 0：

$$W=\sum_{p=1}^{P}\alpha_p d^p X^p,\sum_{p=1}^{P}\alpha_p d^p=0 \tag{9}$$

联合式（8）（9）可得

$$W=\sum_{p=1}^{P}\alpha_p d^p X^p,\sum_{p=1}^{P}\alpha_p d^p=0 \tag{10}$$

根据（8）（10）可得

$$L(W,b,\alpha)=-\frac{1}{2}W^TW+\sum_{p=1}^{P}\alpha_p \tag{11}$$

则有

$$\max W(\alpha)=\sum_{p=1}^{P}\alpha_p-\frac{1}{2}\sum_{p,j=1}^{P}\alpha_p\alpha_j d^p d^j (X^p)^T X^p \tag{12}$$

确定 α 的最优值后，结合（3）（9）式即可得出 W 和 b，此时可得到最优分类判别函数为：

$$\begin{aligned} f(X^p)&=\mathrm{sgn}(W^TX^p+b)\\ &=\mathrm{sgn}(\sum_{p=1}^{P}\alpha_{0p}d^p(X^p)^TX+b_0)\end{aligned} \tag{13}$$

对于线性不可分数据，则将其映射到高维特征向量空间，在映射函数适当且特征空间维数足够高的情况下，大多数非线性可分模式可在特征空间中转换为线性可分模式。

2　支持向量机文本分类模型

2.1　算法实现

结合支持向量机模型，极短文本的处理流程如图 3 所示。

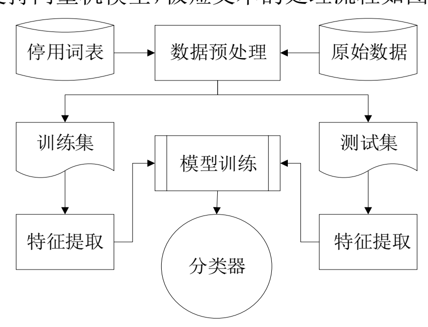


图 3　极短文本分类流程

Fig.3　Extremely short text classification process

经过基本处理的文本信息依然不能被计算机所识别，且其中每个词语对分类贡献不明确。为此，需选用某种方法对其进行特征提取，强化特征词的影响并且减弱非特征词的干扰。

TF-IDF 是一种典型的文本特征提取算法，通过词频与反文档频率的组合计算，有效标志出词语对于分类的贡献。在对分类器进行训练之前，随机将数据按 70%、30%比例分为训练集、测试集。使用训练集对分类器进行训练后，将测试集输入进行验证。

2.2　核函数选择

在构建 SVM 分类模型时，选择一个合适的核函数至关重要。对于内积核函数，常用的有以下四种：

a）线性核函数（Linear）：

$$K(X^p,X)=X^p\times X \tag{14}$$

b）多项式核函数(Poly)：

$$K(X^p,X)=[(X\cdot X^p)+1]^d \tag{15}$$

d 为多项式核函数的最高项的次数。

c）径向基核函数(RBF)：

$$K(X^p,X)=\exp(-\gamma\|X-X^p\|^2) \tag{16}$$

γ 为径向基核函数的参数

d）Sigmoid 核函数(Sigmoid)：

$$K(X^p,X)=\tanh[u(X\cdot X^p)-r] \tag{17}$$

u、r 为 sigmoid 的参数。

四种核函数在不同的应用场景中表现各不相同。在本文特征数远大于样本数的情况下，通常选用线性核函数。

3　实验及结果分析

3.1　模型验证

为了检验模型灵敏度，尤其是预测的准确率以及被错误分类的情况，本文采用了基于混淆矩阵的检验方法。混淆矩阵的结构如下：

混淆矩阵		预测				
		C_1	C_2	…	C_n	合计
实际	C_1	t_{11}	t_{12}	…	t_{1n}	$Cou(C_1)$
	C_2	t_{21}	t_{22}	…	t_{2n}	$Cou(C_2)$
	…	…	…	…	…	…
	C_n	t_{n1}	t_{n2}	…	t_{nn}	$Cou(C_n)$
	合计	PC_1	PC_2	…	PC_n	N

其中，对于 n 阶混淆矩阵

$$T=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n}\\ t_{21} & t_{22} & \cdots & t_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ t_{n1} & t_{n2} & \cdots & t_{nn}\end{bmatrix}$$

当 $i=j$ 时，t_{ii} 表示预测样本中，被正确判别到 C_i 的个数；当 $i\neq j$ 时，t_{ij} 表示本应属于 C_i 类的样本被归为 C_j 类的个数。对模型的检验需要从两个方面来分析，一方面是模型的预测准确度，另一方面是模型处理不同预测时的稳定性。

首先考察模型的准确度。在对抽取的样本容量为 N 的样本进行预测时，正确预测的样本个数占样本总数的比值称为准确率，记作 Ar，即混淆矩阵的迹与样本总数的比值：

$$Ar=\frac{tr(T)}{N}=\frac{\sum_{i=1}^{n}t_{ij}}{N} \tag{18}$$

在相同的环境下，利用程序对抽取的某个样本进行多次预测，每次均可得到一个随机的准确率。在大量实验下，准确率的分布情况如图 4 所示.

存在大量训练样本的条件下，模型的准确率较高。在 100 次实验后，模型的准确率稳定在 98.1%左右。

其次考察模型的稳定性。相对于总预测样本来说，误判的个数越少，则该模型越稳定。本文采用误分度这一概念来刻画模型的稳定程度。误分度的定义如下：

定义 2　记误分度为 Er，则

$$Er = \frac{\prod_{(i,j)} t_{ij}!}{\sum_i^n t_{ij} \times \prod_i^n PC_i!}, i,j = 1,2,\dots,n \tag{19}$$

其中: PC_i 表示第 i 个预测分类中，所预测的个数，即混淆矩阵的列和。

对于误分度，需要从数值和图像两个方面衡量。从数值上看，在 $all(t_{ij} i \neq j) = 0$ 时，

$$Er = \frac{1}{N} \tag{20}$$

在误分情况极少时，误分度趋近于预测样本数的倒数；同时，若特征向量合理，预测样本总数极大，则误分度趋近于 0，符合实际情形。

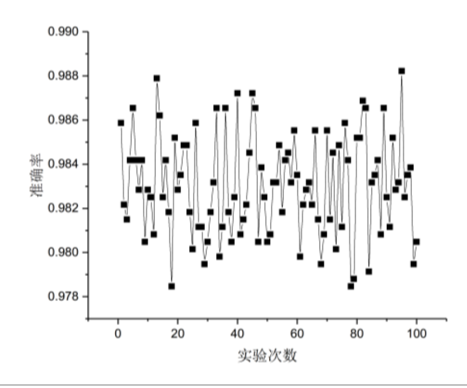


图 4　样本预测准确率

Fig. 4　Sample prediction accuracy

从图 4 看，形成的散点应在某条水平直线附件上下波动。满足以上两点的模型具有较好的稳定性。

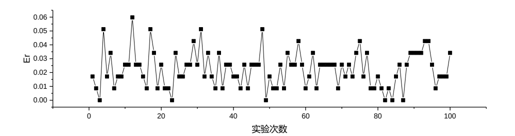
综上，本文称准确率与误分度的共同检验为（1-0）检验模型。在二者均满足各自的检验条件时，能获得较为理想的预测结果。该模型不仅追求较高的成功率，同时还考虑了复杂条件下模型的适用性。在衡量分类优劣时，（1-0）检验模型具有很好的性能。

3.2　基于社管平台数据的实验

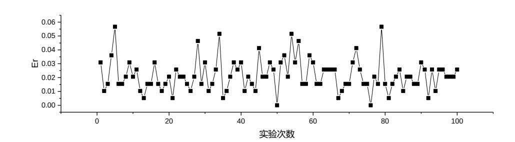
本文以芜湖市“全民社管”软件采集的实际数据，对模型的可用性进行了进一步检验。所谓“全民社管”，即市民发现本市的不文明现象、安全隐患、损坏的公共设施等问题后，通过手机进行爆料，移交政府有关部门处置，从而实现“共建、共治、共享”的社会治理格局。本文的模型即对于每一条上报的案卷，判别该案卷所属类别，以便快速处理。通过该软件，采集到 9906 条极短文本，包含环境卫生、违规广告、施工垃圾、安全隐患、违法占道、公共设施等 6 项内容。将文本输入经过训练后的分类器进行分类，结果如表 1 所示。

为检验模型的误分度，在不改变训练集的前提下，分别选取了测试集中的 600、1000、2000 个数据进行实验，结果如图 5 所示。图 5 显示，在测试集逐渐增大的情况下，误分度有趋近样本数的倒数的趋势；同时，随着样本数的增大，误分度在一条水平直线上波动，该直线略高于样本数的倒数。因此，模型在大规模数据样本的场景下性能较优。

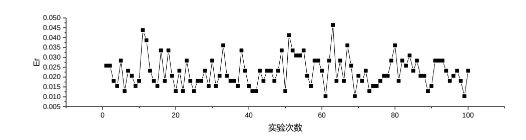
为了验证模型的有效性，进行了对比实验。实验采用 Python 进行数据分析，将数据随机分为 70%的训练集，30%的测试集，对模型进行了 5 次实验，其结果如表 1 所示。在表 1 中，用于评价模型的指标选取了精确度（precision）,召回率（recall）以及 F1 值（f1-score）三种。其中精确度表现了模型对于正样本的区分程度，召回率体现了对负样本的区分程度，而 F1 值则是二者均值。实验数据表明，SVM 相比其他算法有着更高的准确率，且在样本的识别度方面表现较优。



（a）测试样本=600



（b）测试样本=1000



（c）测试样本=2000

图 5　误分度预测

Fig.5　Misclassification prediction

表 1　算法比较实验结果

Table 1　algorithm comparison experiment results

算法选择	评价指标	实验次数 1	2	3	4	5
支持向量机	F1-score	0.98	0.98	0.98	0.98	0.98
	Recall	0.98	0.98	0.98	0.98	0.98
	Precision	0.98	0.98	0.98	0.98	0.98
贝叶斯	F1-score	0.89	0.88	0.88	0.88	0.88
	Recall	0.90	0.89	0.89	0.89	0.89
	Precision	0.91	0.89	0.88	0.88	0.88
决策树	F1-score	0.97	0.97	0.97	0.97	0.97
	Recall	0.97	0.97	0.97	0.97	0.97
	Precision	0.97	0.97	0.97	0.97	0.97
逻辑回归	F1-score	0.95	0.95	0.94	0.95	0.96
	Recall	0.96	0.95	0.95	0.95	0.95
	Precision	0.94	0.95	0.96	0.95	0.96

表 1 表明，支持向量机在精确度、召回率等方面均具有较好表现，分类结果较为理想。

数据集包含环境卫生、违规广告、施工垃圾、安全隐患、违法占道、公共设施等 6 项内容。将文本输入经过训练后的分类器进行分类，分别得到的正确分类数量结果如表 2 所示。

表 2　样本数据分类实验结果

Table 2　Sample data classification experiment results

	样本种类 环境卫生	违规广告	施工垃圾	安全隐患	违法占道	公共设施
样本数	1964	1265	633	217	4359	1594
SVM	1933	1252	630	212	4322	1367
贝叶斯	1854	1123	542	9	4311	996
决策树	1928	1202	626	210	4233	259
逻辑回归	1930	1214	620	206	4302	1170

从表 2 可以看出，支持向量机的分类效果明显优于其他三种算法，且在样本数量较少时表现优异。

在支持向量机的应用中，选择合适的核函数至关重要。现阶段主要有两种选择方案，一是根据前人经验；二是根据

实验结果对比。本文通过对比实验，最终决定选择线性核函数。其实验结果（分类准确率）如表 3 所示。

表 3　核函数比较实验结果

Table 3　Nuclear function comparison experiment results

核函数类型	实验次数				
	1	2	3	4	5
Linear	0.985	0.979	0.985	0.978	0.984
Poly	0.44	0.43	0.43	0.43	0.43
RBF	0.45	0.43	0.43	0.45	0.44
Sigmoid	0.43	0.42	0.44	0.43	0.44

由对比实验可知，线性核函数相比于其他核函数具有极大的优势，在其他核函数分类准确率只有45%左右的情况下，线性核函数仍能保持 98%以上的分类准确率。

4　结束语

本文根据智慧社会管理平台的现实需要，提出了基于贝叶斯决策的极短文本分类模型。模型通过特征词提取保障了关键词的合理性；随后将分类概率与贝叶斯分类器相结合；最后对模型进行检验，实验表明模型具有良好的极短文本分类性能。但模型的误分度稳定性和词语权重合理性等问题仍有待进一步研究。

参考文献：

[1] Butnaru A M, Ionescu R T. From image to text classification: a novel approach based on clustering word embeddings [J]. Procedia Computer Science, 2017：112-120.

[2] Sabbah T, Selamat A, Selamat M H,et al. Modified frequency-based term weighting schemes for text classification [J]. Applied Soft Computing, 2017, 58(9): 193-206.

[3] 刘志康. 一种改进的混合核函数支持向量机文本分类方法 [J]. 工业控制计算机, 2016, 29 (6): 113-114. (Liu Zhikang. An improved mixed kernel function support vector machine text classification method [J]. Industrial Control Computer, 2016, 29 (6): 113-114.)

[4] 李村合, 马敏敏. 增量支持向量机核函数的优化 [J]. 计算机系统应用, 2017, 26 (8): 284-287. (LI Cunhe, MA Minmin. Optimization of kernel function of incremental support vector machine [J]. Computer Systems& Applications, 2017, 26 (8): 284-287.)

[5] 郑立洲. 短文本信息抽取若干技术研究 [D]. 合肥：中国科学技术大学, 2016. (Zheng Lizhou. Research on Several Techniques of Short Text Information Extraction [D].Hefei: University of Science & Technology of China, 2016.)

[6] 任迪, 万健, 殷昱煜, 等. 基于贝叶斯分类的Web服务质量预测方法研究 [J]. 浙江大学学报：工学版, 2017, 51(6): 1242-1251. (Ren Di, Wan Jian, Yin Yuyu,et al. Research on Web service quality prediction method based on Bayesian classification [J]. Journal of Zhejiang University:Engineering Science,2017, 51(6): 1242-1251.)

[7] 池云仙, 赵书良, 罗燕, 等. 基于词频统计规律的文本数据预处理方法 [J]. 计算机科学, 2017,44(10):276-282.(Chi Yunxian, Zhao Shuliang, Luo Yan, *et al*. Text data preprocessing method based on word frequency statistics [J]. Computer Science, 2017,44(10): 276-282.)

[8] 数据堂.停用词集合[DB/OL]. http://www. datatang. com/data/19300/. (Data Hall. Stop word collection [DB/OL]. http://www. datatang. com/data/19300/.)

[9] 孟丹. 基于深度学习的图像分类方法研究 [D].上海：华东师范大学, 2017. (Meng Dan. Research on image classification based on deep learning [D].Shanhai: East China Normal University, 2017.)

[10] Mao X, Zhao G, Sun R. Naive Bayesian algorithm classification model with local attribute weighted based on KNN [C]//Proc of IEEE Information Technology, Networking, Electronic and Automation Control Conference. IEEE, 2017: 904-908.

[11] Mirzaei A, Mohsenzadeh Y, Sheikhzadeh H. Variational relevant sample-feature machine: a fully Bayesian approach for embedded feature selection [J]. Neurocomputing, 2017, 241: 181-190.

[12] 易顺明, 易昊, 周国栋. 采用情感特征向量的 Twitter 情感分类方法研究 [J]. 小型微型计算机系统, 2016, 37(11):2454-2458. (Yi Shunming, Yi Hao, Zhou Guodong. Study on Twitter sentiment classification method based on emotional feature vector [J]. Mini-Micro Systems, 2016, 37 (11): 2454-2458.)

[13] 杨思春, 戴新宇, 陈家骏. 面向开放域问答的问题分类技术研究进展[J].电子学报, 2015,43(8):1627-1636. (Yang Sichun, Dai Xinyu, Chen Jiajun. Research progress of problem classification technology for open domain question and answer [J]. Acta Electronica Sinica, 2015, 43(8): 1627-1636.)

[14] 魏芳芳, 段青玲, 肖晓琰, 等. 基于支持向量机的中文农业文本分类技术研究 [J]. 农业机械学报, 2015, 46 (S1): 174-179. (wei fangfang, duan qingling, xi xiaoxiao, et al. Study on Chinese agricultural text classification technology based on support vector machine[J]. Transactions of the Chinese Society of Agricultural Machinery, 2015, 46 (S1): 174-179.)

[15] 吴家菁, 王杨, 闫小敬,等. 基于 multi-agent 理论的社会网络文体分类方法 [J]. 计算机系统应用, 2014, 23(11):122-126. (Wu Jiajing, Wang Yang, Yan Xiaojing,et al, Chen Fulong. A social network stylistic classification method based on multi-agent theory [J]. Journal of Computer Systems, 2014, 23 (11): 122-126.)